Meeting Summary

The Cancer Genome Atlas's (TCGA) 1st Annual Scientific Symposium: Enabling Cancer Research through TCGA

Gaylord National Hotel
National Harbor, MD

November 17-18, 2011

National Cancer Institute
National Human Genome Research Institute
National Institutes of Health
U.S. Department of Health and Human Services

Table of Contents

Section **Page**

Thursday, November 17

Opening Remarks ……..……...…….. **5**
Lynda Chin, M.D.

Keynote Address and Questions & Answers ……...…….. **5**
Eric S. Lander, Ph.D.

Session I
Chair: Peter W. Laird, Ph.D., M.S.

Lead Talk—The Cancer Epigenome……...…….. **8**
Peter W. Laird, Ph.D., M.S.

Predicting Patient Outcomes with Chained Biological Concept Classifiers …….. **9**
K. James Durbin and Daniel Edward Carlin, M.S.

Lessons Learned from 24 Completely Sequenced AML Genomes …….. **9**
Timothy Ley, M.D.

LRpath Analysis Reveals Common Pathways Dysregulated via DNA Methylation Across Cancer Types ……...……...…….. **10**
Maureen A. Sartor, Ph.D., M.S.

Multi-Cancer Mutual Exclusivity Analysis of Genomic Alterations …….. **11**
Giovanni Ciriello, Ph.D.

Genome-Wide Co-Localization Studies of Somatic Copy Number Alterations and Germline Common Variant Risk Loci in Cancer ……..…….. **12**
Marcin Imielinski, M.D., Ph.D.

Correlating Protein Phosphorylation with Genomic Alterations in Cancer …….. **13**
Jianjiong Gao, Ph.D.

Session II
Chair: Ilya Shmulevich, Ph.D.

Lead Talk: Integrative Analysis and Interactive Exploration of Data from TCGA .. **13**
Ilya Shmulevich, Ph.D.

Absolute Quantification of Somatic DNA Alterations in Cancer Reveals Frequent Genome Doublings in Human Cancers …….. **15**
Scott L. Carter, Ph.D.

Predicting the Impact of Mutations in Cancer using an Integrated Pathway Approach …… 15
Sam Ng

TCGA Computational Histopathology Pipeline Reveals Subtypes and their Molecular Signatures ……...……...…….. 16
Hang Chang and Bahram Parvin, Ph.D., M.S.

Algorithms for Automated Discovery of Mutated Pathways in Cancer ……..…….. 17
Ben Raphael, Ph.D.

The Spectra of Somatic Mutations across Many Tumor Types ….……..…….. 17
Michael S. Lawrence, Ph.D.

An Integrated View into Multivariate Associations Inferred from TCGA Cancer Data ……... 18
Richard Kreisberg, M.S.

Friday, November 18

Session III
Chair: Marco Marra, Ph.D.

Lead Talk: Sequence-Based RNA Profiling: Expression Maps at Base-Pair Resolution ……...…….. 18
Marco Marra, Ph.D.

RetroSeq: A Tool to Discover Somatic Insertion of Retrotransposons …….. ……... 19
Elena Helman

Patient-Specific Pathway Analysis using PARADIGM Identifies Key Activities in Multiple Cancers ……... 20
Josh Stuart, Ph.D.

Morphologic Analysis of Glioblastoma Identifies Morphology-Driven Clusters and Molecular Correlates Associated with Patient Survival ……...…….. 21
Lee Cooper, Ph.D.

Validated Targets Associated with Curatively Treated Advanced Serous Ovarian Carcinoma ……... 22
Douglas A. Levine, M.D.

Massively Parallel Validation of Cancer Mutations and other Variants Identified by Whole Cancer Genome and Exome Sequencing ….……..…….. 22
Georges Natsoulis, Ph.D.

SuperPathway Analyses of Luminal and Basaloid Breast Cancers from The Cancer Genome Atlas Program ……..….. **23**
Christopher Benz, M.D.

Session IV
Chair: David Haussler, Ph.D.

Lead Talk: Large-Scale Cancer Genomics Data Analysis……...…….. **24**
David Haussler, Ph.D.

RF-ACE for Uncovering Non-linear Associations from Heterogeneous Cancer Data …….... **25**
Timo Erkkila, M.S.

Supporting Subtype Characterization through Integrative Visualization of Cancer Genomics Datasets …….……... **25**
Nils Gehlenborg, Ph.D.

Uncovering the Pseudo-Subclonal Structure of Tumor Samples with Copy Number Variation Analysis of Next-Generation Sequencing Data ……...……... **26**
Yi Qiao

Comparison and Validation of Somatic Mutation Callers …….. ……... **26**
Andrey Sivachenko, Ph.D.

Using TCGA Samples to Infer Post-Transcriptional Regulation in Cancer ……... **27**
Pavel P. Sumazin, Ph.D.

Neuroimaging Predictors of Survival, Pathology, and Molecular Profiles in TCGA Glioblastomas …….....……... **28**
David Gutman, M.D., Ph.D.

Closing Remarks …….....……... **29**
Elaine Mardis, Ph.D.

Thursday, November 17

Opening Remarks

Lynda Chin, M.D.; University of Texas, M.D. Anderson Cancer Center

Dr. Chin began by thanking attendees and noting that the goal of this symposium is to showcase some of the outstanding science enabled by TCGA that is being conducted within the consortium and by the larger cancer community. She noted TCGA has built momentum, and its impact is being felt in all areas of basic, translational, and clinical research. This symposium is therefore a venue to share exciting new data and their myriad applications and to convene and build the community of TCGA data users, with the ultimate goals of increasing the uses and accelerating the translation of TCGA data into endpoints that can impact patient outcome. Dr. Chin also noted that this symposium will showcase young investigators in this new area of science, and she thanked members of the meeting committee for their help in screening and selecting talks from the submitted abstracts.

Keynote Address: Cancer Genomics: Fulfilling the Promise

Eric S. Lander, Ph.D.; Broad Institute

Dr. Lander provided an overview of cancer genomics research for the past century, with a focus on where future efforts will lead. In 1914, Boveri proposed that chromosome defects cause cancer. In the ensuing decades, however, this hypothesis was largely replaced with the idea that viruses cause cancer. In 1971, Knudsen proposed the two-hit hypothesis that suggested that viral oncogenes were related to cellular genes. By 1980, the concept that most cancer defects arose from cellular genes had been established, and by 1986, most of the mechanisms causing these defects were known. In 1986, Dulbecco (*Science* 231:1055-1056) noted the importance of sequencing the human genome, and 15 years later, sequencing of the human genome allowed investigators to take systematic approaches to genome studies. Breakthroughs during the first decade of the 21st century (e.g., microarrays, DNA resequencing, RNAi) enabled systematic sequencing of pathways and gene classes implicated in cancer (e.g., BRAF, PIK3CA, EGFR), systematic microarray studies, and integrative genomics. In 2005, the NCI's National Cancer Advisory Board assembled a Working Group on Biomedical Technology that argued that cancer is a genetic, heterogeneous, and understandable disease; a systematic understanding of cancer would thus have major implications in numerous research and healthcare areas. Moreover, the Working Group argued that systematic understanding of caner genome is technologically feasible within the next decade at a modest cost in context.

Although the Human Genome Project was proposed prior to the advent of PCR, the Project caused dramatic changes in the cost of sequencing. As such, a flexible, incremental, evolving plan (The Human Cancer Genome Project) was proposed to investigate genomic loss and amplification, point mutations in coding regions, and chromosomal rearrangements and epigenetic changes. The project set a goal of identifying all genomic alterations significantly associated with all major cancer types by creating a large collection of appropriate, clinically annotated samples and completely characterizing each sample. This proposal engendered pushback in terms of possibility and cost estimates.

In 2006, a pilot project, then called The Cancer Genome Atlas, was launched, and other established genomic organizations such as the Sanger Centre and the International Cancer Genome Consortium (ICGC) became involved. Since the initiation of TCGA, the cost of sequencing has decreased 1000-fold. More than 600 TCGA samples have been analyzed in publications, with more than 1000 whole cancer genomes and more than 9000 cancer whole exomes characterized to date. Dr. Lander noted that such large-scale genomic approaches provide insights into functional classes of genes implicated in cancer, including protein and lipid kinases, lineage survival genes, epigenetic regulators, metabolic enzymes, RNA splicing factors, and translocations. Surprise findings have included identifying *Notch* as an oncogene in T-ALL and a tumor suppressor in squamous skin cancers. Genes identified by TCGA as implicated in carcinogenesis include *BRAF, PIK3CA, EGFR, FGFR2,* and *ALK*.

TCGA aims to create a comprehensive catalog of the cancer genome and to understand how molecular signaling pathways are involved in carcinogenesis. To accomplish these goals, it is essential to identify all driver genes in all cancer types and understand how these genes correlate with clinical phenotype. It is known that many genes are implicated in specific cancers, but only some are significantly mutated. Dr. Lander noted that current lists of "significant genes" are populated yet incomplete. For example, a TCGA analysis of 457 lung cancer specimens identified 843 significant genes, many of which (e.g., the 146 olfactory receptors identified) are not likely to be significant to the cancer process. As such, a significance score must be assigned based on a background model of a constant mutation rate across the genome. However, heterogeneity also creates false positives, making it appear that some genes are mutated at a higher than actual rate. This problem grows more acute as the sample size increases. However, cancer types show distinctive mutation rates and patterns. Mutation rates vary with gene expression; genes that are more highly expressed have lower mutation rates due to transcription-coupled repair. Nonetheless, the regional rate of mutation varies across the genome, and human mutation rates are associated with DNA replication timing for germline and somatic mutations. Late replication explains most of the olfactory receptors and other genes of questionable cancer significance that appear in lists of "significantly mutated" genes. Correcting for variations in the mutation rate removes many of these genes, thus suggesting that learning the mutation rate is a critical concept.

TCGA efforts have supported a wide range of genomic analyses that have led to numerous insights about specific cancers and the global disease. Whole genome analysis has revealed that 10% of the genome is typically involved in focal amplifications and deletions in cancers. However, the search for amplifications is challenged by identifying the driver genes, and homozygosity is necessary for a deletion to be significant. Nearly one quarter of the genome of a given cancer is engaged in arm-level mutations, although the driver genes are not easily identifiable. Methylation analyses have identified subgroup phenotypes of glioma, yet there is no currently available way to identify the associated driver genes. Analysis of translocations has provided evidence of many simultaneously occurring events, yet translocation rates and patterns vary across cancer types. Challenges with regard to translocations include significance and completeness. Integrating events are also observed in many cancers, including ovarian and lung small-cell carcinoma, but large, complete datasets are required to understand these events. Current efforts focus on coding regions, thus potentially overlooking non-genic targets. Microbes have been correlated with many cancers, but proving causation remains a challenge. With regard to understanding germline mutations that explain heritability, large sample sizes are necessary to

determine if such mutations reflect common variants of a small effect or rare variants of a large (greater than or equal to five-fold) effect.

Dr. Lander thus noted that understanding the cancer genome catalog will require integration across tumor types. Thousands of specimens may be required of some cancers to achieve the sufficient sensitivity. To date, TCGA has focused on large, surgically-treatable tumors and pre-treatment tumors. To understand fully the interplay of genomic events in cancer, the initiative must expand its efforts to include assessing intra-tumor diversity and analyzing non-resectable cancers and metastases. Results must then be correlated with clinical information. To this end, Dr. Lander recommended creating a Global Cancer Alliance, a shared knowledge base to which cancer patients can choose to contribute their genomic and clinical data.

Systematic knowledge bases are required to recognize functional pathways and mechanisms in which targets act and to understand cancer vulnerabilities and resistance mechanisms as functions of the cancer genome. Dr. Lander suggested creating a Cancer Therapeutic Roadmap that would enable systematic input of genes into pathways and processes to understand function and provide insight about targets for intervention. Other ongoing efforts, such as the Connectivity Map, a "Google for biologic function," represents a compendium that is connected to the Gene Expression Omnibus, allowing researchers to study expression patterns that result from knocking out specific genes. He noted that the ability to study network effects provides additional data and increases study power, thereby enabling studies of gene function, repurposing of drugs, toxicity studies, interpretation of hits in a screen, and direct small-molecule screening. TCGA is also working with the Cancer Cell Line Encyclopedia, which aims to study complete genomes from 1000 cancer cell lines in which genes have been knocked out systematically. This tool will enable researchers to search for cancers with amplification in a particular gene or genes and assess the downstream effects, thereby facilitating development of resistance catalogs and definition of countermeasures. To meet the overall goal of understanding functional pathways in cancer, the entire cancer community must share open and usable data.

Discussion:

One attendee asked about minimizing redundancy across large-scale efforts. Dr. Lander replied that TCGA gains strength because it is coordinated but not centralized. He noted that scale is key, and projects should begin as pilots to determine their viability and to demonstrate a communal need. He noted that the entire community is collectively responsible for assembling a set of open, high-quality facts. Another attendee asked about systematic efforts to integrate non-coding regions, to which Dr. Lander replied that no systematic efforts on the TCGA scale are currently underway. Many of these regions are small, suggesting that only a few great examples could prove the value of this approach. He noted that the Encyclopedia of DNA Elements (ENCODE) is currently investigating some non-coding regions, although ENCODE and TCGA do not currently integrate their data sets. Another participant asked about targeted therapy, given the heterogeneity of disease. Referencing the Project Achilles, Dr. Lander replied that therapies must be tested in combination or in parallel, even when each drug will likely fail individually. However, there is no infrastructure currently in place to develop and approve multiple agents simultaneously.

Session 1
Chair: Peter W. Laird, Ph.D., M.S.; University of Southern California Epigenome Center

Lead Talk—The Cancer Epigenome
Peter W. Laird, Ph.D., M.S.; University of Southern California Epigenome Center

Dr. Laird began by observing that the torrent of data generated by TCGA exceeds the analysis capacity of TCGA community. DNA methylation alterations in cancer represent one mark that survives processing to naked DNA. CpG dinucleotides are possible targets for methylation that are clustered in CpG islands that tend to be located near promoter regions. In cancer, CpG islands may acquire abnormal focal hypermethylation, and methylated CpG island promoters are transcriptionally silenced in cancer. Moreover, areas of low CpG density may lose DNA methylation in cancer. TCGA efforts have revealed epigenetic silencing of *BRCA1* in serous ovarian cancer (TCGA Research Network. *Nature* 2011;474:609), indicating that germline and somatic mutations are distinct from methylation. Mutated cases appear associated with better survival than epigenetically silenced cases, although it is not clear how these methylated cases respond to PARP inhibitors. TCGA glioma analyses have identified a CpG island methylator phenotype (G-CIMP; Noushmehr, et.al. *Cancer Cell* 2010;17:510) that defines a distinct subgroup of glioma. Four subtypes of glioma (e.g., proneural, mesencyhmal, classical, and neural) have been identified based on expression clusters and methylation clusters. G-CIMP is a subset of proneural GBM with better survival. An almost perfect correlation is observed between mutation in *IDH1* and the G-CIMP phenotype. *IDH1* mutation may cause aberrant CpG island methylation that ultimately allows accumulation of CpG islands. This hypothesis, however, only explains a subset of cases and does not explain G-CIMP $IDH1^{wt}$ cases.

Cross-tumor comparison of methylation among 2275 TCGA specimens and 409 normal tissues reveals that gastrointestinal (GI) cancers cluster. Female-hormone driven malignancies also cluster. Whole genome bisulfite sequencing (WGBS) has been carried out on four TCGA cancers to date, with three lung and three breast tumors in process. WGBS of TCGA tissues indicates that most of the genome is heavily methylated, with methylation-prone regions throughout. Methylation-prone elements are enriched for stem cell polycomb markers, and transcriptional potential is associated with histone H3 methylation. Polycomb target genes in embryonic stem cells are master regulators of differentiation and development that are poised to be turned on during differentiation. Polycomb target DNA methylation begins in normal tissues and becomes exacerbated in cancer. Polycomb crosstalk leads to cumulative stochastic methylation with aging. As such, a transient repressive state becomes a permanently silent state, and the cell loses its ability to differentiate. These cells then become stuck as self-renewing cells that can become target cells with appropriate stimulus. This model would explain the DNA methylation behavior for approximately half of cancer-specifically methylated genes and is consistent with the observation of epigenetic field effects adjacent to tumors. The model also suggests that cancer may start as a differentiation defect consistent with the stem cell-like behavior of cancer cells and with evidence for tumor-initiating cells. The model further suggests that the first steps in oncogenesis may be epigenetic. Therapeutic cloning strategies that use human embryonic stem cells or induced pluripotent stem cells should incorporate screening for PRC2 DNA methylation abnormalities.

Methylation-prone CpG islands (Berman, et.al. *Nat Genet* 2012;44:40-46) are regions where methylation encroaches into CpG islands in cancer cells. Regions of focal hypermethylation and long-range hypomethylation coincide, and a subset of the cancer epigenome has partially lost methylation. These hypomethylated "oceans" correspond to lamin attachment domains in the nuclear periphery, indicating that the epigenome has a spatial organization.

In summary, Dr. Laird noted that a CpG island methylator phenotype has been identified in glioblastoma that is likely related to *IDH1* mutation. Polycomb repressor binding in embryonic stem cells predisposes DNA methylation to cancer; this polycomb repressor predisposition is seen across cancer types. Focal hypermethylation and long-range hypomethylation coincide in partially-methylated domains (PMDs), and epigenetically unstable PMDs are associated with nuclear lamina attachment and late-replicating regions.

Predicting Patient Outcomes with Chained Biological Concept Classifiers
K. James Durbin and Daniel Edward Carlin, M.S.; University of California, Santa Cruz

Dr. Carlin began by noting that cell-specific programs observed in stem cells are aberrantly activated in many types of cancer cells. As such, his laboratory is working to develop a method to detect these programs. Classifier chaining uses stem cell expression to build a signature of "stemness" that can then be used to classify cancers. In this approach, cancer expression datasets are mapped to a standard set of genes and normalized by quantile. Once a robust stem cell classifier was identified, applied learning was used to mine expression data from breast, colorectal, glioblastoma, lung, and ovarian cancers from TCGA for stemness levels and their relationships to cancer subtypes. This approach can be used to learn on one type of cancer and then be applied to other types. For example, this approach was trained using 80% of BRCA breast cancer data to discern signatures for basal versus luminal subtypes and then applied to the remaining 20% of BRCA data and to all TCGA ovarian cancer data. Future efforts will focus on obtaining gene signatures from a binary classifier. Such signatures can recognize two-way distinctions. A cancer sample is run on each classifier individually, thereby defining a signature for each cell based on its performance in each classifier. Results from these experiments can reveal where the cell resides in the hierarchy of development as well as its tissue of origin. Each time that binary classifiers are evaluated, the one that performs best is retained. While this approach overfits to a specific dataset, a wealth of biologic data is captured.

Lessons Learned from 24 Completely Sequenced Acute Myeloid Leukemia (AML) Genomes
Timothy Ley, M.D.; Washington University in St. Louis

Dr. Ley began by stating the little is known about the key initiating mutations for most patients with AML, excepting canonical translocations. However, AML tumor tissue can be conveniently and repeatedly accessed, and most samples are relatively free of contaminating normal cells. Furthermore, many AML genomes are diploid, and low-resolution genomic screening (cytogenetics) is an established paradigm for classifying disease and supporting treatment decisions. Favorable-risk cases can be treated lightly upfront, whereas adverse-risk cases require transplant. However, two-thirds of cases are intermediate risk, revealing a need for biomarkers that can further stratify disease. Moreover, each AML genome contains hundreds of mutations, and all of the mutations are present in all tumor cells. This finding suggests that all mutations could have arisen simultaneously or that clonal evolution has generated hundreds of relevant

mutations per genome. Given the improbability of these hypotheses, a more plausible model involves hematopoietic stem cells, which accumulate an average of 14 mutations per year until an AML-initiating mutation is acquired. Once the critical cell has been created, it undergoes clonal expansion, thus explaining why all AML mutations have the same read-count frequency when sampled deeply by NGS data. Sequencing the AML genome thus captures the relevant mutations plus all others that have accumulated in the cell. But how many mutations are required to cause AML? A comparison between the mutational burdens seen in M3 AML (initiated by *PML-RARA*) versus M1 AML with normal karyotype (NK) should indicate the same total number of mutations in both genomes, most of which will be random and irrelevant. However, M1 should also feature some novel mutations that represent tumor initiation, whereas mutations shared between M1 and M3 genomes will indicate disease progression.

To determine the distribution of recurrent mutations, 24 AML tumor/normal genome pairs were sequenced. These efforts identified 10,597 validated somatic mutations, including 21 recurrently mutated genes (ten in M1 only; one in M3 only; and eleven common to both). All *de novo* AMLs have founding clones, and some have subclones. All four cohesion complex genes are mutated in AML.

The TCGA AML200 project includes whole-genome sequencing of 50 cases of *de novo* AML tumor/normal pairs, 150 exomes (also tumor/normal pairs), 173 transcriptomes, and 192 methylation arrays. Cases were chosen to represent all known subtypes of AML in terms of karyotype and morphology. Age is a predictor of the number of mutated genes in AML. RNASeq has been used to identify gene fusions from *de novo* assembly of AML transcriptomes. Deep digital sequencing shows how AML undergoes clonal evolution at relapse. Clonal behavior leading to relapse will be critical for therapy, as founding clones must be removed to cure the disease. Dr. Ley noted that these data provide a lower bound on the number of AML mutations necessary to cause the disease. Several genes likely to be important have been identified, although it is difficult to tell whether these will pan out until many hundreds of AML genomes are sequenced, and recurrence is assessed. One attendee asked if parallel *FLT3*-like mutations were observed in patients without *FLT3* mutations, to which Dr. Ley replied that they were generally not found in AML genomes, although other cooperating mutations do clearly exist.

LRpath Analysis Reveals Common Pathways Dysregulated via DNA Methylation across Cancer Types
Maureen A. Sartor, Ph.D., M.S.; University of Michigan

Dr. Sartor noted that the relative contribution of epigenetic mechanisms to carcinogenesis is poorly understood--do epigenetic mechanisms target genes and pathways similarly to somatic mutations? The Illumina HumanMethyaltion27 Bead Chip platform assesses the percentage methylation of more than 27,000 CpG sites across the genome, and several studies have been published testing for genes with aberrant methylation in their promoter regions. Interestingly, most of these publicly available datasets study cancer. As such, Dr. Sartor noted that the time is ripe for an integrative analysis using data from TCGA and NCBI's Gene Expression Omnibus (GEO) to test whether certain pathways or gene groups are commonly dysregulated across cancer groups via DNA methylation during the cancer pathogenesis. This approach analyzed pathways using the interactive logistical regression-based LRPath platform (http://lrpath.ncibi.org; Sartor MA, et.al. *Bioinformatics* 2009;25:211-217). Advantages of

LRPath include strong performance for datasets with large and small sample sizes, the ability to test directional and non-directional tests, the ability to interpret random sets without the need for significance values to be "approximately normally distributed," generation of identical significance values for repeated runs (e.g., no dependence on permutations), and a flat *p*-value distribution under the null.

Dr. Sartor noted that LRPath can be used to test many different types of gene sets, including those identified from pathway analysis, literature, experimental results, and targeted approaches. LRPath also provides clustering options and filtering capability. The Illumina 27 Bead Chip assesses the percent methylation for more than 27,000 sites, spanning more than 14,000 genes. This platform was used to study pathways commonly altered as a result of DNA methylation in tumor versus normal tissues for ten cancer types, each of which featured at least 1000 sites with a greater than ten-fold change in methylation. Hypomethylated pathways identified included those related to immune response (e.g., chemokine and cytokine activity, responses to stimulus and inflammation, receptor binding activities, and peptidase activities) and epidermal development. Hypermethylated pathways included those associated with nerve development, embryonic development, homeobox signaling, sequence-specific DNA binding, and voltage-gated potassium channels. Some gene sets were selectively methylated in breast or prostate cancers. In summary, Dr. Sartor noted that pathways affected by differential methylation were surprisingly concordant across cancer types. DNA repair, one of the most commonly affected pathways in cancer development, is depleted in differentially methylated genes. For most tumor types, a change in CpG methylation within a pathway affected similar genes. An integrated analysis of biologic concepts dysregulated via methylation across ten cancer types identified concepts that were affected in multiple cancer types that support biologically important findings. A subset of known cancer pathways appears to be commonly dysregulated via DNA methylation across cancers. She noted that this approach has yet to be applied to investigate hypermethylation in cancer subtypes. One attendee asked how "normal" was defined at the epigenetic/proteomic level, to which Dr. Sartor replied that, while cancer-specific methylation probes tend to be consistent across a wide range of normal tissues, "normal" means that the cell of origin has a profile similar to that of the tumor under study. Another participant asked if gene sets in close proximity were affected by the same event of hypomethylation. Dr. Sartor replied that immune response genes are close together, so it is possible that proximal genes are affected similarly.

Multi-Cancer Mutual Exclusivity Analysis of Genomic Alterations
Giovanni Ciriello, Ph.D.; Memorial Sloan-Kettering Cancer Center

Dr. Ciriello began by observing that recurrent genomic alterations target specific pathways and that functional alterations that target the same pathway frequently occur in a mutually exclusive manner. The Ciriello laboratory has developed the Mutual Exclusivity Modules (MEMo) tool (Ciriello G, et.al. *Genome Res* 2011;Oct 12:Epub ahead of print) to systematically identify mutually exclusive alterations that target oncogenic pathways across multiple cancer types. MEMo has been applied to five TCGA projects (glioblastoma, ovarian, colorectal, uterine, and breast cancers). Mutually exclusive patterns of alteration have been identified in several oncogenic pathways, including Rb signaling, p53 signaling, DNA repair, and PI(3)K/Akt. MEMo did not find PI(3)K/Akt modules in ovarian cancer, although multiple low-frequency events target the PI(3)K pathway. 463 samples of invasive breast cancer have also been examined. The heterogeneity of the disease suggests that basal and luminal subtypes may

actually represent distinct diseases. These analyses addressed whether pathways are differentially modified on the basis of these subtypes and whether the PI(3)K pathway is altered by other means in basal tumors. *PTEN* is down regulated in basal tumors independently of copy number status, and *PTEN* downregulation activates *Akt* phosphorylation. *AKT3* is over-expressed in basal breast cancer; 30% of basal tumors have some alteration in this pathway. In summary, Dr. Ciriello noted that PI(3)K/Akt signaling is consistently altered in cancers, albeit to different extents and by different mechanisms. Mutual exclusivity analysis across multiple cancers unveils the underlying heterogeneity of the disease, thus suggesting subtype-specific candidate therapeutic targets.

Genome-Wide Co-Localization Studies of Somatic Copy Number Alterations and Germline Common Variant Risk Loci in Cancer
Marcin Imielinski, M.D., Ph.D.; Broad Institute

Dr. Imielinski began by stating that germline risk variants and somatic mutations are two central facets of cancer genomics. Common cancers are 2-4-fold more likely in first-degree family members of affected patients, although cancer risk is mediated by complex polygenic inheritance. Rare, highly-penetrant variants can explain 5% and 20% of the heritable risk for breast and colorectal cancers, respectively. However, genome-wide association studies (GWAS) have identified common, mildly-penetrant variants that explain 10%, 23%, and 6% of the heritable risk for breast, prostate, and colorectal cancers, respectively. Germline cancer susceptibility loci frequently mutated in cancer include *TP53*, *APC*, *RB1*, and *CDKN2A*. GWAS has been used to quantify the overlap with somatic copy number alteration studies (SCNA; Beroukhim R, et.al. *Nature* 2010;463:899-905) and to determine the significance against a null model built via permutation. 297 cancer loci from the National Human Genome Research Institute's GWAS database were compared against significant SCNA peak regions from pan-tumor and sub-analyses of 20 tumor types. Analysis of cancer versus non-cancer-associated GWAS regions showed a significantly enriched overlap with respect to regions of amplification but not with regions of deletion. Dr. Imielinski then asked whether germline SNP status confers risk for specific somatic alterations. Allelic bias has been observed in somatic copy number alterations, and an allelic distortion test can measure whether a given allele has unique features that contribute to bias. This test measures the frequency a given allele is amplified or deleted at each heterozygous SNP. None of 36 cancer-GWAS loci that intersected a somatic amplification peak region showed significant allelic distortion. However, distortion was observed at the *CCND1* locus across many cancers and cell lines; the "C" allele was amplified much more frequently than the "T" allele. The biologic significance of this phenomenon is not known. In summary, Dr. Imielinski noted that significant overlap of germline GWAS peaks and SCNAs (e.g., amplifications, and amplifications plus deletions) was observed across cancer types, providing the first evidence for genome-wide colocalization of germline susceptibility variants and somatically altered loci.

One participant asked why heterozygous loci were chosen for these experiments, to which Dr. Imielinski replied that results were cleaner because a simple statistical test can be used to determine selective advantage. These experiments focused solely on somatic mutations.

Correlating Protein Phosphorylation with Genomic Alterations in Cancer
Jianjiong Gao, Ph.D.; Memorial Sloan-Kettering Cancer Center

Dr. Gao began by noting that RPPA acts as a quantitative, high-throughput, multiplexed, and inexpensive ELISA in which each slide is developed using a single antibody. RPPA data have been generated for six tumor types (breast, ovarian, colorectal, endometrioid, glioblastoma, and kidney) using protein antibodies that include PTEN, TP53, ER, and AR. Antibodies have been developed to phosphoproteins and proteins implicated in numerous signaling pathways, including Rb, p53, and PI(3)K/Akt. Dr. Gao noted that ERBB2 mRNA, protein, and phosphorylation levels are well correlated in breast cancer, whereas ER and GATA3 protein levels differ in breast cancer subtypes. *PTEN* deletion and under-expression are correlated with elevated phospho-AKT (pAKT) in glioblastoma and breast, ovarian, and colorectal cancers. PhosphoAKT has diverse targets that regulate proliferation, invasion, and apoptosis; it contributes to breast cancer progression and confers resistance to conventional therapies. But what genomic event activates AKT in breast tumors? Loss of *PTEN*, but not *PTEN* mutation, is strongly associated with pAKT, whereas *PIK3CA* mutations and *RTK* amplifications are not associated with elevated pAKT. To assess other genomic events that could explain AKT phosphorylation in breast cancer, an enrichment test was applied using all GISTIC ROIs and frequently mutated genes. Preliminary results suggest that *CAMK1D* amplification and *AKT1 E17K* mutation may play roles. In summary, Dr. Gao noted that genomic and proteomic data correlate well at the level of individual genes and proteins. However, downstream effects are more difficult to link to genomic events. Some (but not all) cases of AKT phosphorylation can be explained, and analysis of genomic event combinations may be helpful in this regard. Systematic analysis of all antibodies is needed, and correlations among protein data may help to elucidate active cancer signaling pathways.

Session II
Chair: Ilya Shmulevich, Ph.D.; Institute for Systems Biology

Lead Talk: Integrative Analysis and Interactive Exploration of Data from TCGA
Ilya Shmulevich, Ph.D.; Institute for Systems Biology

Dr. Shmulevich began by noting that TCGA is producing a wealth of rich and heterogeneous data (e.g., gene expression, CN, methylation, clinical data) that are continuous, discrete, categorical, and, in some cases, missing. As such, analyzing these data in an integrated fashion is computationally and statistically challenging. Pairwise analysis seeks to identify correlations between two data elements, such as gene expression and methylation or mutation and clinical outcome, using either single or different data types. To identify associations, integrated analysis requires a feature matrix that can incorporate clinical information, tumor characteristics, and other data that may be generated externally using other algorithms. Such analyses have identified many interesting associations in TCGA data. For example, in gliomas, *IDH1* status has been shown to be related to the CpG island methylator phenotype (Noushmehr H, et.al. *Cancer Cell* 2010;17:510-522). As an example of a categorical association with a continuous variable, elevated expression of *ESR1* has been identified as a distinguishing feature of the luminal subtype of breast cancer (Sorlie T, et.al. *PNAS* 2003;100:8418-8423). Integrated analyses of TCGA data have also shown that most mutations of *TP53* occur in the DNA-binding domain and that samples where this is true exhibit lower expression levels of downstream target expression.

In ovarian carcinoma, *RAB25* expression is associated with promoter methylation (TCGA Research Network. *Nature* 2011;474:609-615); high methylation levels correlate with lower expression. Moreover, *PRAC* expression and methylation are both strongly correlated with the clinical parameter, "anatomic organ subdivision."

Dr. Shmulevich noted that one goal of integrative analysis is to understand mechanistically how disruption causes molecular networks to cease functioning. For example, *Wnt* signaling is aberrant in colorectal cancer, and integrative analysis of RNA transport, methylation, and translation data can provide insight into the mechanisms that affect transcription and function in this cancer type. Such analyses of colorectal cancer data have identified many relationships between clinical features. For example, *MLH1* methylation is associated with microsatellite instability (MSI) and CIMPH methylation, whereas *BRAF* mutations associate with MSI/CIMP expression clustering. Based on parameters such as histologic type, lymphatic invasion, and tumor stage, integrative analysis can identify an "aggressiveness summary" that details the correlation between genes and clusters of clinical features that associate with aggressive disease (House CD, et.al. *Cancer Res* 200; 70;6957-6967). Interactive tools can identify clinically-associated genomic hotspots.

However, relationships between gene expression and cancer subtype or outcome are often not one-to-one, especially when data are heterogeneous. For example, genes A and B, when taken individually, may not be associated with outcome C, but their combination may provide a predictive value for outcome C. To help with multivariate analysis of heterogeneous data, Random Forest (RF)-ACE (http://code.google.com/p/rf-ace) has been developed. RF-ACE is a multivariate statistical inference method based on ensembles of decision trees that seeks to reveal significant associations between features in the input data matrix. RF-ACE has a high predictive power and is resistant to over-fitting. It can handle mixed variable types, does not require imputation of missing variables, and is based on random sub-sampling rather than on combinatorial search. Statistical testing removes redundant features, enabling a fast and portable implementation in C++. RF-ACE can address computational challenges such as mixed data types, tens of thousands of features in hundreds of samples, missing data, correlated features, and nonlinear, noisy, and multivariate relationships. This method selects particular features of interest, and data are split into smaller disjoint sets, essentially building "decision trees" that can be aggregated to improve identification of important features and to observe how these features behave in comparison to artificial contrasts. An importance score is assigned to each feature, as certain features are strongly associated with numerous other features ("bundles"). While this tool shows the features associated with each chromosome, the snapshot is static. An interactive tool, Regulome Explorer (explorer.cancerregulome.org), has been developed that allows users to explore multivariate relationships in the data (e.g., the association of *IDH1* mutations with the C-GIMP methylation phenotype in GBM; *TP53* mutations associated with *CDKN2A* and *TP73* methylation). However, these tools must be integrated with information from the literature, protein-protein interactions, and other databases to get a complete understanding of the associations implicated in cancer processes. To look at associations in conjunction with the literature, Pubcrawl can be used to reveal the similarity between two terms in PubMed. Dr. Shmulevich noted that this tool will be integrated into Regulome Explorer in the near future.

One attendee noted that some features of aggressiveness are correlated, yet a Fisher's test assumes that all variables are independent. Dr. Shmulevich replied that RF can handle correlated

features, although this tool was not used when analyzing the aggressiveness data. Because features are not independent, a weighted Fisher's method was used to "tamp down" features that are so heavily associated as to overwhelm the signal from lesser-associated features. He noted that Regulome Explorer allows users to filter results according to importance scores and correlation to help weed out false discoveries.

Absolute Quantification of Somatic DNA Alterations in Cancer Reveals Frequent Genome Doublings in Human Cancers
Scott L. Carter, Ph.D.; Broad Institute

Dr. Carter began by noting that purity and ploidy determine the power to detect mutations in cancer genomes; the observed copy number (CN) signal is proportional to locus concentration for sequencing and hybridization methods. Sequencing can be used to identify subclonal point mutations, although discrete allelic fractions are obscured by tumor purity and local copy number. The recently-developed ABSOLUTE algorithm uses cellular multiplicity (an integral allele count) to classify point mutations. Equivalent nucleotide substitution frequencies for clonal and subclonal point mutations rule out contamination, thereby providing a multiplicity estimate to classify mutated genes. ABSOLUTE was applied to search for frequently homozygous genes, such as tumor suppressors, in TCGA ovarian cancer specimens. These analyses revealed that *TP53* was present at two or more copies per cancer cell, suggesting that its mutation is likely to be an early event in ovarian carcinogenesis. Analysis of genome doublings indicates a bimodal distribution of ploidy in human cancer. Absolute allelic CN data suggest that high-ploidy samples evolved via a genome-doubling event. Genome doubling varies across human cancers; at least 60% of ovarian cancers analyzed have at least one doubling, whereas ALL had almost none. In a pattern observed across many cancer types, genome doubling occurs after aneuploidy, whereas loss of heterozygosity (LOH) precedes doubling. Genome-doubled tumors have more copy alterations, yet genome-doubled ovarian cancer evolves differently. Of the 15 *NF1* mutations identified in a set of 214 TCGA ovarian carcinomas, 13 occurred in non-doubled samples, in which case they were homozygous. Dr. Carter noted that these results show that selection acts specifically on recessive inactivation of *NF1*. Moreover, no amplified mutations in *NF1* were observed in doubled samples. In contrast to *p53*, *NF1* mutators do not progress via genome doubling. Clinical correlations associated with genome doubling include patient age at diagnosis and time to recurrence. One attendee asked why the mutant allele spectrum differs when samples are contaminated, to which Dr. Carter replied that these samples tend to contain an excess of germline variants.

Predicting the Impact of Mutations in Cancer using an Integrated Pathway Approach
Sam Ng; University of California, Santa Cruz

Dr. Ng began by stating that his research aims to identify driver mutations and their modes of action, such as gain- and loss-of-function (GOF and LOF). Many recurrent, low-frequency mutations remain poorly characterized, and understanding the modes of action of these mutations can provide insight into disease mechanisms and enable treatment. The Ng laboratory has recently developed PARADIGM, a method that utilizes functional genomic data and pathways to predict LOF or GOF. LOF and GOF occur in the context of pathways, and PARADIGM uses CN expression upstream and downstream of a mutation and sets of pathways to infer gene-level activities. Discrepancy scores between upstream and downstream data differ between mutated

and non-mutated samples; a more negative discrepancy score is indicative of LOF. Given the same network topology, how likely would a gain or loss of function be called? Passenger mutations are not discrepant. Applying discrepancy analysis to TCGA colorectal cancer specimens identified four non-discrepant genes with sufficient pathway annotations. In summary, Dr. Ng noted that discrepancy analysis combines functional genomic data such as CN and expression with pathway information to differentiate between neutral, GOF, and LOF mutations. To date, the approach has successfully identified *RB1* LOF in glioblastoma and *NFE212* GOF in lung cancer. Discrepancy analysis is specific and does not identify discrepancies concordant with MutSig calls. Identifying potential GOFs can reveal possible treatments that could apply to sensitive tumors or cell lines.

Discussion:

One attendee asked if the method could be extended to identify a switch of function. Dr. Ng replied that this may be possible, although it would likely be difficult. Dr. Ng noted that PARADIGM collects and collapses all mutations that occur on one gene. One attendee suggested that it may be useful to consider that different amino acid changes on one gene may have opposing effects. Another participant asked why a bimodal distribution of discrepancies is observed in mutant samples, to which Dr. Ng replied that this could reflect a neutral mutation, even if it is non-silent.

TCGA Computational Histopathology Pipeline Reveals Subtypes and their Molecular Signatures

Hang Chang and Bahram Parvin, Ph.D., M.S.; Lawrence Berkeley National Laboratory

Dr. Parvin began by stating that the computational histopathology pipeline captures the molecular basis of each morphometric subtype. In the case of GBM, specimens are curated by removing tissue sections that contain artifacts (e.g., tissue folds, pen marks, scanning anomalies). Analysis of a typical GBM specimen requires one week of computing time, produces large datasets, and must account for biologic variation. The Parvin laboratory has developed robust and efficient image analysis algorithms (tcga.lbl.gov) to compute morphometric features and meta-features. These tools enable subtyping based on selected features or reduced dimensionality and facilitate the association of molecular information with morphometric subtypes. The algorithm enhances nuclear segmentation in the presence of technical variations. New images are normalized against reference images to construct a probability model based on a Gaussian distribution. Seed detection methods provide the shape and local statistics. Once a nuclear feature can be delineated, a representation must be completed to identify structural features. Based on cellularity and nuclear size at the patient level, four GBM subtypes have been identified. This algorithm enables the user to calculate the ability of each subtype to predict survival. Since a tumor is heterogeneous, can it be queried for subtypes at the block level to learn about tumor composition? Based on tumor histology, GBM patients can be classified as belonging to one subtype, and high cellularity and low nuclear size are better predictive of a more aggressive therapy. In conclusion, Dr. Parvin noted that many approaches can be used to parse genomic data, and different indices lead to alternative subtyping, which in turn enable alternative biological interpretations. As such, genomic association has the potential to reveal new insight.

One attendee asked if the Parvin group had compared the heterogeneity index with SNP data, to which Dr. Parvin replied that this was forthcoming. Morphologic features have been correlated only with gene expression data at this point.

Algorithms for Automated Discovery of Mutated Pathways in Cancer
Ben Raphael, Ph.D.; Brown University

Dr. Raphael noted that one key challenge for cancer genome sequencing is distinguishing between driver and passenger mutations observed in high-throughput analyses. Recurrent mutations—those occurring more frequently across a patient cohort than would be expected by chance--can serve as one rubric to prioritize mutations as possible drivers. To identify recurrent mutations, a statistical test can be applied for each individual gene, followed by a multiple hypothesis correcting procedure to adjust for the number of statistical tests performed. The first two TCGA publications used such a procedure but identified a relatively small list of statistically significant genes. Moreover, many genes were mutated in multiple patients at levels insufficient to achieve statistical significance. Because cancer is a disease of pathways, the standard approach to identify driver mutations begins with networks of known pathways. Limitations of this approach include testing only existing pathways without considering pathway topology and challenges resolving issues of crosstalk between pathways. Dr. Raphael then discussed two methods less biased by prior knowledge, HotNet and Dendrix, which were developed to examine large numbers of genes. HotNet uses a predefined network to identify connected subnetworks mutated in a significant number of patients. This method acknowledges the importance of local network topology (Vandin F, et.al. *J Comp Biol* 2011;18:507-522). When applied to 316 TCGA ovarian cancer specimens, HotNet identified 27 subnetworks that contained seven or more genes. Twelve of the 27 subnetworks showed significant overlap with known pathways or protein complexes. In contrast to HotNet, the Dendrix algorithm supposes that driver mutations are rare and that a cancer pathway has one driver mutation per patient, thereby imposing mutual exclusivity between mutations. The algorithm also assumes that pathways should have high coverage and seeks to identify genes with high coverage. When applied to 199 AML specimens, Dendrix identified two (non-validated) top-scoring sets with many co-occurrences within and between the sets. On these same specimens, HotNet identified five (non-validated) subnetworks containing five or more genes. HotNet analysis of 514 TCGA breast cancer specimens identified 13 (non-validated) networks containing eight or more genes. Dr. Raphael concluded by noting that future efforts will focus on incorporating additional data types such as gene expression and methylation analyses and performing pre/post filtering of predictions. One participant asked if there were any known exceptions to the exclusivity assumption, to which Dr. Raphael replied that many gene sets are not exclusive, while others are exclusive but do not interact with the network. Another attendee asked what happens if the driver mutation actually occurs prior to activation of *p53* or another gene identified using these methods. Dr. Raphael replied that these methods do not provide temporal information.

The Spectra of Somatic Mutations across Many Tumor Types
Michael S. Lawrence, Ph.D.; Broad Institute

Dr. Lawrence began by noting that the sequencing of thousands of genomes from multiple tumor types has revealed vast differences in the rates, prevalence, and types of mutations. The MutSig scoring algorithm has been developed to identify significantly mutated genes, assuming a

background mutation rate that is uniform across sequence contexts, patients, and genes. However, the mutation rate is heterogeneous across genes, thus challenging efforts to identify driver genes. Two distributions with the same average number of mutations could be quite different from one another. Analysis of data from 457 TCGA lung cancer specimens has identified 843 significantly mutated genes, many of which (e.g., olfactory receptors) are highly unlikely to be driver genes for lung cancers. Highly expressed genes have lower mutation rates (Chapman MA, et.al. *Nature* 2011;471:467-472). Given that the background mutation rate and replication rate vary greatly across the genome, late replication explains most of the noted olfactory receptors. Dr. Lawrence noted that the mutational landscape is not "flat" based on gene expression level, replication time, and mutation rate. After filtering on the basis of influence on "neighbor" genes, the 843 genes previously identified reduce to 52 genes. MutSig can then be run on a pan-cancer set to identify the top significantly mutated genes across tumor types.

One attendee noted that there is a relationship between mutation count, the probability of a mutation to arise, and the proliferative advantage a mutation confers. Dr. Lawrence replied that this relationship is assumed, although it may prove false. The MutSig algorithm will soon enter beta testing and will be made available to the community.

An Integrated View into Multivariate Associations Inferred from TCGA Cancer Data
Richard Kreisberg, M.S.; Institute for Systems Biology

Mr. Kreisberg began by noting that visual analytics tools identify advantageous ways to explore large, heterogeneous datasets to formulate testable hypotheses about different aspects of the data. He noted that a successful tool is interactive, allows the user to reason easily based on what he/she sees, and connects data to other sources and tools. Emergent technologies include browser-based graphical rendering (e.g., SVG, Canvas, WebGL, and Flash), scalable cloud services, adaptive data models (e.g., noSQL technologies), graph databases (e.g., Neo4J, OrientDB), and graph computation to reason on the topology of the data. The open-source Regulome Explorer (http://explorere.cancerregulome.org) tool incorporates applications such as Random Forest, colorectal aggressiveness scoring, and All Pairs Significance to assemble the association topology. Future technologies include network topologies that provide explicit, retrievable states and allow the user to import data. One attendee asked if Regulome Explorer users can load data into the program. Mr. Kreisberg noted that this would be possible by modifying the code, although the algorithm is not set up explicitly to enable this function.

Friday, November 18

Session III

Chair: Marco Marra, Ph.D.; British Columbia Cancer Agency

Lead Talk: Sequence-Based RNA Profiling: Expression Maps at Base-Pair Resolution
Marco Marra, Ph.D.; British Columbia Cancer Agency

Dr. Marra began by noting that RNA sequencing (RNASeq) enables analyses of gene and isoform expression and detection of gene fusions, expressed mutations, and cancer subtypes. In addition, miRNA sequencing enables analyses of cancer subtypes and regulatory networks. These techniques thus offer the opportunity to create a map of the cancer genome, and RNASeq

data can be used to define cancer types and subtypes. A cohort of RNASeq tools (Garber M, et.al. *Nat Methods* 2011;8:469-477) is evolving along with a wealth of RNASeq datasets. Dr. Marra noted that gene discovery efforts rose rapidly with the increase in the number of reads, followed by a plateau. Exon wiring maps represent the primary strength of RNASeq data, as transcripts may be altered in numerous ways during splicing to affect protein products. Expression profiling data provide examples where exons are dropped or retained. *De novo* assembly can be used to identify structures not found using alignment-based platforms. Platforms such as Trans-ABySS (Robertson G, et.al. *Nat Methods* 2010;7:909-912) align contigs back to the genome, thus enabling alignment-independent detection of gene fusions, alternative transcripts, internal and partial tandem duplications, and insertions/deletions. Trans-ABySS has been applied exhaustively to verify 39 AML gene fusions, 25 of which are novel. This pipeline requires manual interrogation but produces high verification rates.

RNASeq has recently been used to verify mutations previously defined by WES in TCGA squamous cell lung tumors. When combined with genome data, RNASeq data can rapidly verify mutations and confirm fusions detected using low-pass sequencing of colorectal cancer specimens. TCGA is also currently building and storing miRNA sequencing data, with approximately 3000 miRNA Seq profiles representing 18 cancer types currently stored at TCGA's Data Coordinating Center. In animals, miRNAs may be generated through several pathways. miRNA biogenesis produces mature miRNA and miRNA*, and non-canonical miRNA variants (e.g., isomiRs) may further expand the target gene repertoire. 191 miRNA libraries have been sequenced using TCGA AML specimens, revealing 270-422 known and 16 unknown miRNAs. These data reveal the ratio of star- and mature-strand miRNA. miRNA sequencing may also be used to cluster cancer subtypes, and mRNA and miRNA data may concord. RNASeq may also be used to inform about anti-sense gene expression and its potential regulatory consequences. Antisense transcription regulates TRα alternative splicing, which is associated with epigenetic splicing. RNASeq can also address antisense-correlated splicing. Strand-specific RNASeq provides knowledge of individual strands to which sequences map (Parkhomchuk D, *Nucleic Acids Res* 2009;37:e123; Levin JZ, et.al. *Nat Methods* 2010;7:709-715).

One attendee inquired whether verification of a mutation identified by RNASeq requires a specific nucleotide. Dr. Marra replied that RNASeq data are typically matched against genomic data; if they concord at a particular locus, then the locus is verified. Another participant commented that using RNASeq for screening generates a high rate of false positives for samples that do not have the known mutation. Dr. Marra responded by noting that a mutation identified from RNASeq data requires normal tissue to conclude that it is a somatic event. However, if no normal specimen is available, RNASeq can be used for expressed recurrent mutations, which may not be somatic events.

RetroSeq: A Tool to Discover Somatic Insertion of Retrotransposons
Elena Helman; Broad Institute

Ms. Helman began by noting that retrotransposons are mobile genomic elements that copy and paste themselves across the genome via an RNA intermediate, thereby producing two copies of the original element. Retrotransposons occur naturally in cancer and are drivers of genome evolution. While they comprise more than 40% of the human genome, most retrotransposons are

inactive, although some remain "hot." Retrotransposons represent a major source of genetic variation, spanning approximately 10,000 polymorphic sites. It has been estimated that two European individuals differ by 600-1000 retrotransposons. Two of the most abundant retrotransposon elements are LINE-1 (L1), which contains two open reading frames, and ALU, which relies on the L1 retrotransposition machinery. Retrotransposon insertions into the genome can disrupt protein function, affect promoters, create or disrupt sites for RNA splicing, and lead to further genomic rearrangement. Aberrant retrotransposon insertions have been identified in breast, lung, and colorectal cancers. To identify the extent of somatic retrotransposon insertions throughout the cancer genome using paired-end sequencing data, the Meyerson laboratory has developed RetroSeq, a tool that identifies the positions of putative retrotransposon insertions by aligning reads to the retrotransposon consensus sequence and locating clusters of pair-mates. The tumor and normal genomes are then compared to identify sites unique to the cancer genome.

RetroSeq was shown to be sensitive and specific in a simulation that searched for 226 L1 and 732 ALUs inserted into a BAM file. L1 insertions were then examined in nine TCGA colorectal cancer tumor/normal pairs to identify 1470 L1 germline events and 221 L1 somatic events. Future studies in this area will include experimental validation of RetroSeq (in progress), extension to other tumor types, and integration of orthogonal data. In conclusion, Ms. Helman noted that RetroSeq leverages paired-end sequencing data to localize somatic retrotransposon insertions. This approach, which identifies novel insertions present in a tumor but not in its matched normal tissue, has provided evidence for reactivation of retrotransposon mobilization in cancer. She noted that the approach uses paired reads in which only one end is unique but does not consider reads in which both ends align to the repeat.

Patient-Specific Pathway Analysis using PARADIGM Identifies Key Activities in Multiple Cancers

Josh Stuart, Ph.D.; University of California, Santa Cruz

Dr. Stuart began by observing that the comparison of multiple data types can be overwhelming. In cellular systems, some of the underlying machinery is known due to data collection and curation efforts such as Reactome, KEGG, and Pathway Commons. However, data integration is the key to interpreting gene function correctly. Gene expression does not always indicate activity; downstream effects often provide clues. For example, if a highly expressed transcription factor is turning on targets, then one could infer that the factor is involved. However, if the transcription factor is expressed only at a low level, different conclusions may be reached, and multiple modalities may be required for certainty. The Stuart laboratory has recently developed PARADIGM (Vaske CJ, et.al. *Bioinformatics* 2010;26:i237-i245), an integrative approach based on probabilistic models that provides detailed models of gene expression and interaction. This approach analyzes data from a cohort of patients using a pathway model of cancer to determine whether a gene of interest is active. For ovarian cancer, *FOXM1* emerges; it is central to cross-talk between DNA repair and cell proliferation. *IPL* genes stratify ovarian cancer by survival time. *MYC* is characteristically altered in colorectal cancer. Differential subnetworks thus identified can form "super-pathways" that can then identify master regulators that could predict drug response. For instance, PARADIGM analysis of TCGA breast cancer specimens predicts that *PLK1* is a target for treating basal breast cancer. HDAC inhibitors are similarly identified as luminal hub markers, as the HDAC network is down-regulated in basal breast cancer cell lines, and basal breast cancers are resistant to HDAC inhibitors. PARADIGM also supports

discrepancy analysis, which examines differences when the pathway analysis is run twice (once with upstream neighbors and again with downstream neighbors). Pathway discrepancy provides an orthogonal view of the importance of mutations by enabling the probing of infrequent events and supporting the detection of non-coding mutation impacts and the presence of pathway compensation. PARADIGM also enables users to associate the presence of a mutation with potential pathway activities, thus inferring a connection between a mutation and phenotypes. Pathway analyses can also inform a global "pan-cancer map" by connecting molecular subtypes across tissues.

In summary, Dr. Stuart noted that information flow should be structured to accurately model gene activity using multi-modal data, with an initial focus on known cancer biology. Current efforts seek to identify new genes and interactions and to stratify patients into pathway-based subtypes. Sub-networks are predictive markers and can be used to simulate scenarios such as drug inhibition, and even rare mutations can be assessed for their biologic significance. PARADIGM has recently been included in Firehose, thus enabling public access to CPU-intensive results.

Morphologic Analysis of Glioblastoma Identifies Morphology-Driven Clusters and Molecular Correlates Associated with Patient Survival
Lee Cooper, Ph.D.; Emory University

Dr. Cooper noted that his group focuses on imaging technologies and is currently using whole-slide imaging of TCGA specimens to correlate scans of frozen tissue with data from molecular studies. TCGA provides scans of frozen tissues and diagnostic-block permanent sections, which are analyzed at 20X magnification. The Cooper laboratory also conducts pathology evaluations (e.g., percentage necrosis, percentage tumor nuclei) on these specimens. He noted that glioblastoma is heterogeneous in appearance and contains cells representing numerous morphologies (e.g., astro- and oligo-type cells). Current efforts are underway to create algorithms that describe clusters of GBM morphology and to assess whether patients cluster on the basis of morphology. Image analysis captures the whole-slide image, for which morphology signatures can be calculated, followed by correlative analysis. A morphology "engine" circles each nucleus in a cell slide, enabling each cell to be described individually on the basis of its characteristics. A patient morphology profile is then created from the composite cell data and fed into a clustering engine. Once the patient cluster labels have been established, they are fed into a correlative engine to determine whether the morphology clusters correlate with recognized genetic alterations. Clustering analysis of 200 million nuclei from162 TCGA GBM specimens has identified three prognostically-significant morphologic groups named for the functions of associated genes (e.g., cell cycle, chromatin modification, and protein biosynthesis). Efforts are currently underway to devise representative nuclei for each group that will aid in visualization. This method was validated in a separate set of 84 GBM specimens from the Henry Ford Hospital, where survival trends correlated with data from TCGA. At present, no definitive association has been made between survival and molecular subtype/pathology, although the nuclear lumen localization was most highly enriched in cluster-associated genes. In conclusion, Dr. Cooper noted that whole-slide images contain signals related to molecular status and outcome. Image analysis can provide scalable, quantitative measurements of cellular morphology. Datasets such as those generated by TCGA present unique opportunities to

correlate morphology with genomics and patient outcome, although more complex models will be required to account for disease heterogeneity.

One attendee asked Dr. Cooper to elaborate on the relationship between tumor nuclei and stromal components. He noted that no single feature is significant for prognosis, although clustered features provide some prognostic significance. It was also noted that TCGA has now updated its requirement for imaging magnification from 20X to 40X.

Validated Targets Associated with Curatively Treated Advanced Serous Ovarian Carcinoma
Douglas A. Levine, M.D.; Memorial Sloan-Kettering Cancer Center

Dr. Levine noted that cancer genomics supports discovery and technology development, whereas applied cancer genomics can address clinical questions. TCGA is well-suited to support the latter goal due to its richly annotated data. He stated that treatment of ovarian cancer begins with an operation to remove all visible cancer, followed by chemotherapy. Survival rates from this approach can be short-term (five months) or long-term (more than five years). However, a subset of patients who have advanced ovarian cancer show no evidence of recurrence after one set of treatments (initial surgery plus chemotherapy). Differences in gene expression profiles between these patients and long-term survivors whose cancer has recurred will shed light on mechanisms of drug resistance in these cancers. A recent study by the Levine laboratory included patients with advanced-stage high-grade serous ovarian cancer who underwent primary cytoreductive surgery and platinum-based chemotherapy. Data were available for 14 patients from MSKCC and 16 non-overlapping TCGA patients, all with Stages III or IV ovarian cancer, who were curatively treated. 42 patients who recurred and survived more than five years were available as controls for each of these datasets. NanoString gene expression used to validate 86 targets from TCGA and MSKCC identified 19 overlapping genes. This approach was validated externally using 57 independent fresh-frozen, paraffin-embedded tissues. These analyses showed that CYP4B1 was overexpressed in curatively treated serous ovarian cancer patients. Dr. Levine noted that these studies excluded platinum-resistant patients; thus, long-term survival may reflect a taxane-based phenomenon.

Massively Parallel Validation of Cancer Mutations and other Variants Identified by Whole Cancer Genome and Exome Sequencing
Georges Natsoulis, Ph.D.; Stanford University

Dr. Natsoulis noted that most genomic projects begin with a large number of potential variants, which ultimately becomes narrowed. Two methods, OS-Seq and single-strand circularization, address multiple objectives. OS-Seq synthesizes capture probes on the flow-cell lawn (Myllykangas S, et.al. *Nat Biotech* 2011;29:1024-1027) and captures the target region from cancer genomes. Two primer probes are used to provide "double strand" coverage of the target, thereby improving mutation discovery based on both strands. OS-Seq, which analyzes fresh DNA from flash-frozen tissues, can also target loci-like extended exons and provide even coverage of genomic region targets. The tool's advantages include higher specificity and sensitivity of mutation detection, higher accuracy by targeting of any non-repetitive human genome region, accurate variant discovery, identification of rearrangement breakpoint sequences, efficient workflows, and low (< 1 μg DNA) sample requirements.

Single-strand genomic circularization features a single-stranded substrate compatible with FFPE material and capture probes that can be placed at any point in the fragmented DNA. A pilot study assessed the approach's yield and specificity of detection in 628 genomic regions from high-quality genomic DNA from flash-frozen tissue and low-quality DNA from FFPE tissue matched by organ and individual. Dr. Natsoulis noted that FFPE processing can introduce artifacts at the rate of approximately one per 10-15 kB (one error per five genes). Results of the analysis showed 85% heterozygote detection over a 120 kB target region. Specific classes of artifacts observed included transitions and transversions. The two methods can be applied in tandem to carry out single-lane mass validation of WGS. A pilot project using WGS and WES of matched normal blood, primary gastric tumor, and ovarian metastasis identified 386 coding variants that included SNVs, in/dels, and SVs (see the open-access OligoGenome Resource (http://oligogenome.stanford.edu; Newburger DE, et.al. *Nucleic Acids Res* 2011;Nov 18:[Epub ahead of print]). Future efforts for OS-Seq include validating mutations and rearrangements from cancer genomes and conducting "onconome" and exome applications. Single-strand circularization efforts aim to carry out follow-up clinical applications using archival FFPE specimens. Both methods are scalable for single-lane validation of cancer genomic projects. Dr. Natsoulis noted that these methods have numerous applications beyond targeting, including analysis of cDNA.

SuperPathway Analyses of Luminal and Basaloid Breast Cancers from TCGA
Christopher Benz, M.D.; Buck Institute for Research on Aging

Dr. Benz began by noting that clinical subtypes of breast cancers inform prediction, whereas intrinsic subtypes inform prognosis. The biological and clinical heterogeneity of breast cancer is most evident by its different intrinsic transcriptome subtypes (e.g., PAM50), which has yet to be employed clinically. With the exception of the HER2 subtype, the pathways and signaling networks driving and distinguishing other transcriptome subtypes (e.g., basaloid, luminal A and luminal B) remain largely undefined. To discover pathway differences between these three intrinsic subtypes, the pathway inference tool, PARADIGM, was used to analyze approximately 500 TCGA breast cancer samples. PARADIGM integrates DNA copy number and transcriptome data to infer patient-specific pathways from multi-dimensional cancer genomics data (Vaske, et.al. *Bioinformatics* 2010;26:1237-1245). The algorithm generates a heatmap to show clustering of pathway activities into potential networks. Annotation of genes that are *c-MYC* repressed and *c-MYC* activated indicated differentially-activated subnetworks within the consensus clusters. Pathway enrichment identified four common differences shared by luminal A and B that differentiated these subtypes from basal cancers. For example, the FOXA1/ER pathway is higher in luminal cancers, whereas the HIF1A pathway is elevated in basal cancers. The two luminal subtypes also show differential network hub activities, with major differentiating hubs including MYC/Max, FOXM1, PLK1, and MYB. Analysis of TCGA dataset shows that luminal B is associated with a worse overall survival than luminal A. These analyses indicate that activity hubs can distinguish overall survival; cohort dichotomization by the MYC/Max or FOXM1 pathway activities prognostically defines outcome as effectively as does luminal status. Dr. Benz concluded by noting that unsupervised consensus clustering based on PARADIGM inferred activities produced by pathway activity clusters with significant intrinsic breast cancer subtype associations. Pairwise comparisons between intrinsic breast cancer subtypes identified elevated FOXA1/ER and lower HIF1A pathway activities as shared features differentiating luminal and basal cases. SuperPathway analyses identified MYC/Max, FOXM1, MYB and PLK1 as network

hubs with elevated activities in luminal B versus luminal A breast cancers, with two of these showing comparable prognostic value and survival associations as luminal status. SuperPathway analyses may help identify pathway and signaling differences between clinical and intrinsic breast cancer subtypes and ultimately point to subtype-specific therapeutic strategies.

One attendee asked if subsets of luminal B cancers relate to progesterone receptor (PR) proliferation. Dr. Benz replied that the subsets are related to proliferation, although not of PR. He noted that PARADIGM currently contains approximately 1300 curated pathways and that SuperPathway analysis identifies features with ten or more interconnections.

Session IV
Chair: David Haussler, Ph.D.; University of California, Santa Cruz

Lead Talk: Large-Scale Cancer Genomics Data Analysis
David Haussler, Ph.D.; University of California, Santa Cruz

Dr. Haussler began by describing the Cancer Genome Hub (CGHub) currently being constructed to store the massive volume of data (e.g., BAM and VCF files) being produced by TCGA, the NCI's Therapeutically Applicable Research to Generate Effective Treatments (TARGET), and Cancer Genome Anatomy Project/Cancer Genome Characterization Initiative projects. The CGHub is designed for 25,000 cases with an average of 200 gigabytes per case. Thus, five petabytes of data are expected to be deposited, although the resource can be scaled to hold 20 petabytes and can be co-located with other hosting facilities. CGHub goals include enabling the direct comparison and combined data analysis of many large-scale cancer genomics datasets, aggregating sufficient data to provide the statistical power to attack the full complexity of cancer mutations, setting standards for data storage and exchange, encouraging data sharing, and maintaining compatibility (e.g., data formats and access coordination) with other large-scale genomics efforts, including EGA, dbGaP, ICGC, the 1000 Genomes Project, and ENCODE.

Dr. Haussler noted that, given the same BAM files, different mutation calling pipelines do not completely agree. Researchers are just beginning to investigate accuracy and consistency in the detection of structural variation using a case study between UCSC and the Broad Institute to analyze whole-genome GBM data from TCGA. Of the 18 genomes available, sixteen were used to analyze tumor and normal for GBM at both centers. The BamBam calling pipeline used at UCSC detected 167 gene fusions in these analyses, whereas the dRanger program used at the Broad Institute detected 188. Of the detected fusions; 136 are potentially overlapping events. Two tumor samples contained a majority of the top-ranked called events. Whole-genome sequence information showed that independent events lead to somatic homozygous loss of the tumor suppressors *CDK2NA/B*. In 11 of the 16 cases analyzed at both sites, similar events led to homozygous loss of *CDKN2A/B*. Analysis of normal cases is underway. Dr. Haussler noted that chromothripsis has been observed in one GBM specimen. This complex fusion creates multiple mutations and is difficult to analyze. *EGFR* amplification/mutation was also commonly reported; 11 of 17 samples featured chromosome 7 amplifications that included *EGFR*. Four of these eleven samples also featured *EGFRviii* mutations. Exon 2-7 deletion was observed at low copy, suggesting that it likely occurred after amplification events. Copy number states, which are generated by plotting the overall CN against the minority allele CN, could perhaps be used to identify driver mutations. However, GBM tumors are not purely clonal and contain an arbitrary

fraction of subclone parts. Dr. Haussler suggested that a tumor can be thought of as an ecosystem in which subclones compete; the more aggressive subclones are the more successful. System complexity may be preserved through cooperativity or homeostasis.

One attendee commented that break points occur in repetitive regions, which are not included in these analyses. Dr. Haussler noted that sophisticated algorithms can be used with global CN analysis to help identify some of these regions. Another participant asked if the subclones observed on the chromosome 6p arm could represent non-tumor cells, to which Dr. Haussler replied that these analyses account for the total amount of infiltration of non-tumor cells.

RF-ACE for Uncovering Non-linear Associations from Heterogeneous Cancer Data
Timo Erkkila, M.S.; Institute for Systems Biology

Mr. Erkkila began by observing that an annotated feature matrix that integrates various data types can uncover associations from the data. An annotated feature matrix may include hundreds of samples per cancer types, and variables may be heterogeneous (e.g., categorical, numerical, binary). To address this issue, the Random Forest (RF) algorithm has been created to select features from heterogeneous data. RF supports mixed-type data and missing values, and predicted targets may represent any type. Moreover, data transformations are unnecessary, and RF supports multivariate and nonlinear associations. However, the algorithm has a non-normalized importance score that merely ranks associations, and prediction performance can still be enhanced. An RF implementation, Random Forest with Artificial Contrast Ensembles (RF-ACE; http://code.google.com/p/rf-ace), provides the added flexibility to support string literals and various data formats and to interface easily with default parameter options. RF-ACE requires a normalized importance score and includes a statistical testing framework and improved predictive power from the Gradient Boosting Tree (GBT) algorithm.

A pilot study was carried out using RF-ACE to identify 19 significant associations for *PRAC* in data from colon and rectal cancer specimens. The top three of these "core features"--*HOXB13*, anatomic organ subdivision, and promoter methylation—were subjected to the GBT algorithm to build predictors for novel or missing data. A pipeline was developed to integrate data resources and analyze aggregate data with RF-ACE. All associations for all cancer types were stored in a database, and the Regulome Explorer browser (http://explorer.cancerregulome.org) was used to explore further these associations. In summary, Mr. Erkkila noted that RF-ACE combines useful features from various established algorithms for a generic and rapid implementation that is well suited to analyze TCGA data. Novel aspects include *p*-values assigned for associations and GBT for prediction. One participant commented that outputs from various sources contain redundancies, to which Mr. Erkkila replied that compact sets of good predictors are being established. However, this redundancy reduction is part of a roadmap plan for algorithm development that has yet to be implemented.

Supporting Subtype Characterization through Integrative Visualization of Cancer Genomics Datasets
Nils Gehlenborg, Ph.D.; Harvard Medical School

Dr. Gehlenborg began by noting that the identification of tumor subtypes has multiple implications for early detection, prevention, and treatment of cancers. Analyses of TCGA GBM

specimens have revealed four GBM subtypes (e.g., proneural, neural, classical, and mesencyhmal) that can be characterized by abnormalities in a small set of genes that include *PDGFRA*, *EGFR*, *NF1*, and *IDH1*. Data used to classify patients according to subtype represent a variety of analyses, including mRNA, copy number, and mutation status. These results were based on VisBricks, a recently-developed multiform visualization method for large sets of heterogeneous data (Lex A, et.al. *IEEE Transactions on Visualization and Computer Graphics* 2011;17:2291-2300). TCGA gene-level GBM data were subjected to GenePattern consensus non-negative matrix factorization and GISTIC 2.0 analyses. Different numbers of clusters in mRNA data were identified, and other groupings of data types (e.g., CN, miRNA, methylation) were added to see how patients travel among the groupings. The visualization tool, implemented in the Caleydo Visualization Framework, displays clusters as individual, moveable columns that can also integrate heat maps. Other data types, such as clinical data, external classifications, multivariate data, and batch information, can also be included and mapped onto pathways or as Kaplan-Meier plots. The visualization tool can also load data and results from analysis pipelines such as the Broad Institute's Firehose.

Uncovering the Pseudo-Subclonal Structure of Tumor Samples with Copy Number Variation Analysis of Next-Generation Sequencing Data
Yi Qiao; Boston College

Dr. Qiao began by noting that tumor samples will always contain a mixture of DNA from mutated and normal cells. The degree of contamination in the tumor by normal tissue can be calculated by the ratio read depths for tumor and normal in sequencing data (e.g., BAM files) using copy number analysis. Analysis of individual chromosomes reveals differences in tumor purity that cluster because the tumor contains more than one subclone or features several biologic structures that could contribute to the observed model (e.g., aggregate mutations from cancer stem cells). A biologically-motivated, CNV caller-independent model has been developed that can simultaneously estimate the normal cell admixture ratio and tumor heterogeneity, thereby offering a rapid decomposition of subclone structure that can be applied prior to downstream analysis (e.g., SNP calling). Future directions for developing this model include validation and working with capture data. One attendee asked how the model represents uncertainty, to which Dr. Qiao replied that it may be unable to differentiate based solely on CN.

Comparison and Validation of Somatic Mutation Callers
Andrey Sivachenko, Ph.D.; Broad Institute

Dr. Sivachenko began by noting that SNVs are defined simply as single nucleotide differences from a reference tissue. In an ideal situation, SNVs would be identified by resequencing and reading out the results. However, SNVs can be hard to call due to multiple issues with library preparation, and sequencing and data processing can produce a spectrum of SNV-like events. Issues that may confound the analyses include alignment quality around the event, strandness or orientation of supporting reads, sufficient coverage, and sequence context. A successful approach must protect against two types of false positives, those identified when there is no actual event (e.g., from misread bases, a sequencing artifact, or a misaligned read) or a germline event due to low coverage in normal tissue. Cross-sample comparison was initiated with a reference set of TCGA specimens to compare, evaluate, and improve mutation-calling algorithms. Comparison alone allows the user only to contrast the callers against each other, thus ultimately necessitating

a validation process. These studies used data from Phase III of TCGA (20 lung squamous cell carcinoma specimens that were subjected to WES at the Broad Institute). The same sequencing data were called at four centers (Washington University, The Broad Institute, UCSC, and the Baylor College of Medicine) using different algorithms, and resulting call sets were shared between the centers for comparison. RNASeq was used for validation. These experiments identified shared versus center-specific events, noting that, while there was a large overlap, many calls were made by only one center. Center-specific calls generally had different properties that could represent specific false-positive modes of each caller or a specific strength of a given center's pipeline. These comparisons revealed a tendency to call center-specific events at coverages different from those where the shared events are located. Some center-specific calls were questionable upon "manual review," although many were convincing. RNASeq was used as a validation set that offers an independent library construction and different protocol with the same sequencing technology. It is possible to call mutations *de novo* from aligned RNASeq data, although this approach may be too conservative. If it is assumed that *de novo* DNA Seq mutation is sufficiently conservative, weaker evidence from RNASeq than what would be required for a stand-alone *de novo* call can be considered as validation. Dr. Sivachenko noted that sensitivity depends on coverage and allelic fraction. Because original calls feature a range of allelic fractions, it is generally unwise to ask for a fixed, low number of observations in RNASeq. However, the allelic fraction strongly correlates between RNASeq and DNA Seq.

When looking for SNVs in RNASeq, every called mutation site with coverage in RNASeq above a certain threshold can be considered as "covered." If a covered site has at least two reads with an alternate allele as evidenced by RNASeq, it can be considered to be "validated." In conclusion, Dr. Sivachenko noted that a framework has been established within TCGA for evaluating and improving mutation calling algorithms. Efforts are ongoing to validate mutations using RNASeq as validation. Currently, none of the four centers has made its caller publicly available, although it is expected that centers will do so soon. It was also noted that low-coverage in DNA can be problematic; some genes cannot be assessed by RNASeq, thus adding bias to the analyses.

Using TCGA Samples to Infer Post-Transcriptional Regulation in Cancer
Pavel P. Sumazin, Ph.D.; Columbia University

Dr. Sumazin described mechanisms that post-transcriptionally regulate microRNA (miRNA) expression and activity. miRNAs are known to act as tumor suppressors and as oncogenes. Dr. Sumazin stated that it is important to understand transcriptional and post-transcriptional regulation of miRNAs, regulation by miRNAs, and regulation of miRNA activity to understand how miRNAs and genes interact. Large-scale, same-sample profiles of mRNA and miRNA expression provide the data necessary for computational predictions. Tight post-transcriptional control of miRNA biogenesis leads to significant swings in mature miRNA expression. The processing machinery for these events includes canonical biogenesis regulators and non-canonical regulators. For example, the newly-predicted non-canonical regulator, *DDX10*, upregulates miRNA biogenesis in the GBM cell line, SNB19. miRNome-wide profiling in response to regulator silencing suggests that there is an abundance of miRNA-specific regulation in GBM cell lines, and miRNAs respond differently to the silencing of different regulators.

miRNA-activity regulators represent two types: sponge regulators that compete for miRNA programs that regulate other RNAs, and non-sponge regulators that activate or suppress miRISC-mediated regulation of target RNAs. Regulation between *PTEN* and *PTENP1* is an example of sponge regulation. *PTEN* is a tumor suppressor and a key regulator of cancer. *PTEN* and *PTENP1* have common miRNA regulators; changes to *PTENP1* expression modify the post-transcriptional regulatory program that targets *PTEN*. TNRC6 proteins are required for miRISC function and are examples for non-sponge miR-activity regulators. Deleterious somatic mutations to *TNRC6A* may contribute to tumorigenesis of gastric and colorectal cancers. Using genome-wide screening for modulators, Dr. Sumazin and his colleagues identified a miRNA program-mediated regulatory (mPR) network that includes a sub-network of co-regulating GBM drivers. These established drivers of gliomagenesis form a tightly regulated mPR sub-network. Moreover, *PTEN* regulates tumor cell growth rates, and changes to *PTEN* expression correlate with glioblastoma cell growth rates. Silencing of *PTEN* mPR regulators, many of which have never been implicated in cancer and whose locus is deleted in samples where *PTEN* is intact, accelerated tumor cell growth in the same manner as silencing *PTEN*. These results suggest that deletions at various loci can cooperatively regulate distal tumor suppressors and oncogenes through post-transcriptional interactions. These findings provide a new direction for identifying driver mutations by integrating DNA-sequencing analysis with regulatory networks.

Neuroimaging Predictors of Survival, Pathology, and Molecular Profiles in TCGA Glioblastomas

David Gutman, M.D., Ph.D.; Emory University

Dr. Gutman began by noting that his research group studies features that can be derived from radiologic imaging. He stated that GBM is the most common form of primary brain tumor and has a 14-month median survival. The Emory University In Silico Center develops human and/or machine-based assessments of image features, such as the standardized imaging feature, VASARI. A VASARI feature set consists of 30 features (e.g., percent necrosis, proportion enhancing tumor) that describe the size, location, and appearance of a magnetic resonance imaging (MRI) image set. MRI images provide a global view of a tumor, with the caveat that a small tumor adjacent to the motor area has a vastly different outcome than a small tumor located in the frontal lobe. To assess tumor imaging properties systematically, data from 72 GBM patients were obtained from the Cancer Imaging Archive. The contrast enhancing the tumor was an imaging-based predictor of survival; the percent of contrast enhancement was significantly associated with shorter survival. Moreover, the mesenchymal GBM subtype was noted to have significantly lower rates of non-contrast enhancement compared to other GBM subtypes, whereas the proneural subtype was associated with a low degree of contrast enhancement. The images revealed that *EGFR*-mutant GBMs (11/49 patients) were larger than wild-type *EGFR* GBMs, whereas *TP53*-mutant GBMs (9/49 patients) were smaller than those that were wild-type. Dr. Gutman concluded by observing that imaging-based features can provide important prognostic information even after accounting for other clinical variables. Current qualitative work suggests that genotypes may be associated with imaging phenotypes. Future work includes increasing the sample size, moving from an ordinal to a continuous-based assessment of tumor compartments (e.g., volumetrics), and including more sophisticated feature extraction to include texture/size/location and voxel-based assessments.

Closing Remarks
Elaine Mardis, Ph.D.; Washington University School of Medicine

Dr. Mardis thanked participants for their enthusiasm and interest in TCGA. She noted that the original notion of TCGA was established in 2005, and the project has grown exponentially since its earliest incarnation, both in terms of the number of tumor types and the multitude of data being processed in TCGA pipelines. TCGA production pipeline follows on those established for other large-scale genomic projects such as the Human Genome Project, HapMap, and 1000 Genomes. As of November 2011, thousands of tumor/normal pairs are in TCGA analysis pipeline. TCGA tissue specimens are quality-checked for pathology and DNA/RNA, and those that pass these QC steps are subjected to a number of analytical platforms. The goal is to achieve a 360-degree integration and mining of the petabytes of available data. In 2012, TCGA expects to publish results from analyses of colorectal, AML, breast, endometrial, kidney clear cell, lung adenocarcinoma, and head and neck cancers. In addition to whole exome sequencing, whole genome sequencing is also underway in multiple tumor types, revealing a multitude of genomic alterations and illuminating the impact of structural variation on genomics. TCGA enterprise is also being expanded to include pilot projects on FFPE-preserved tissues and mouse models of human cancers, projects to study rare tumor types, and efforts to integrate TCGA efforts with those from ICGC and the NCI's Clinical Proteomic Tumor Analysis Consortium (CPTAC), which will carry out proteomic analyses of colorectal, breast, and ovarian cancer specimens from TCGA. TCGA has also established a committee to assess studies in mouse models that will focus on prostate cancer, melanoma, non-small cell lung cancer, and three breast cancer models. She thanked attendees for attending the symposium, noting that feedback is welcome in preparation for next year's symposium.

The meeting was then adjourned.